Big Data and Modernizing Federal Statistics: Update

Bill Bostic
Associate Director
Economic Programs Directorate

Ron Jarmin Ph.D.
Assistant Director,
Research and Methodology Directorate

April 16, 2015

Census Bureau “Big Data” Research Agenda

- Methodological
- Computational
- Policy / Legal
- User and Stakeholder Engagement

Update on Efforts

- New Center in R&M (need name)
  - Hub for Bureau efforts in this area
  - Lead projects
  - Affiliated projects managed in other directorates
  - Searching for Chief

Big Data Center Projects

- Innovation Measurement Initiative (IMI)
- MIT Workshops
  - Big Data and Commodity Flows (joint with the Bureau of Transportation Statistics)
  - Big Data and Privacy
  - Big Data and Adaptive Survey Design
- Big Data Class
- Sandbox

Innovation Measurement Initiative

- Collaborative research project between Census, University of Michigan, Ohio State and University of Chicago
- Integrate university data on federally funded research grants with Census Bureau data assets
- Produce statistics consistent with the Bureau's economic and social measurement mission and directly relevant to the data provider.

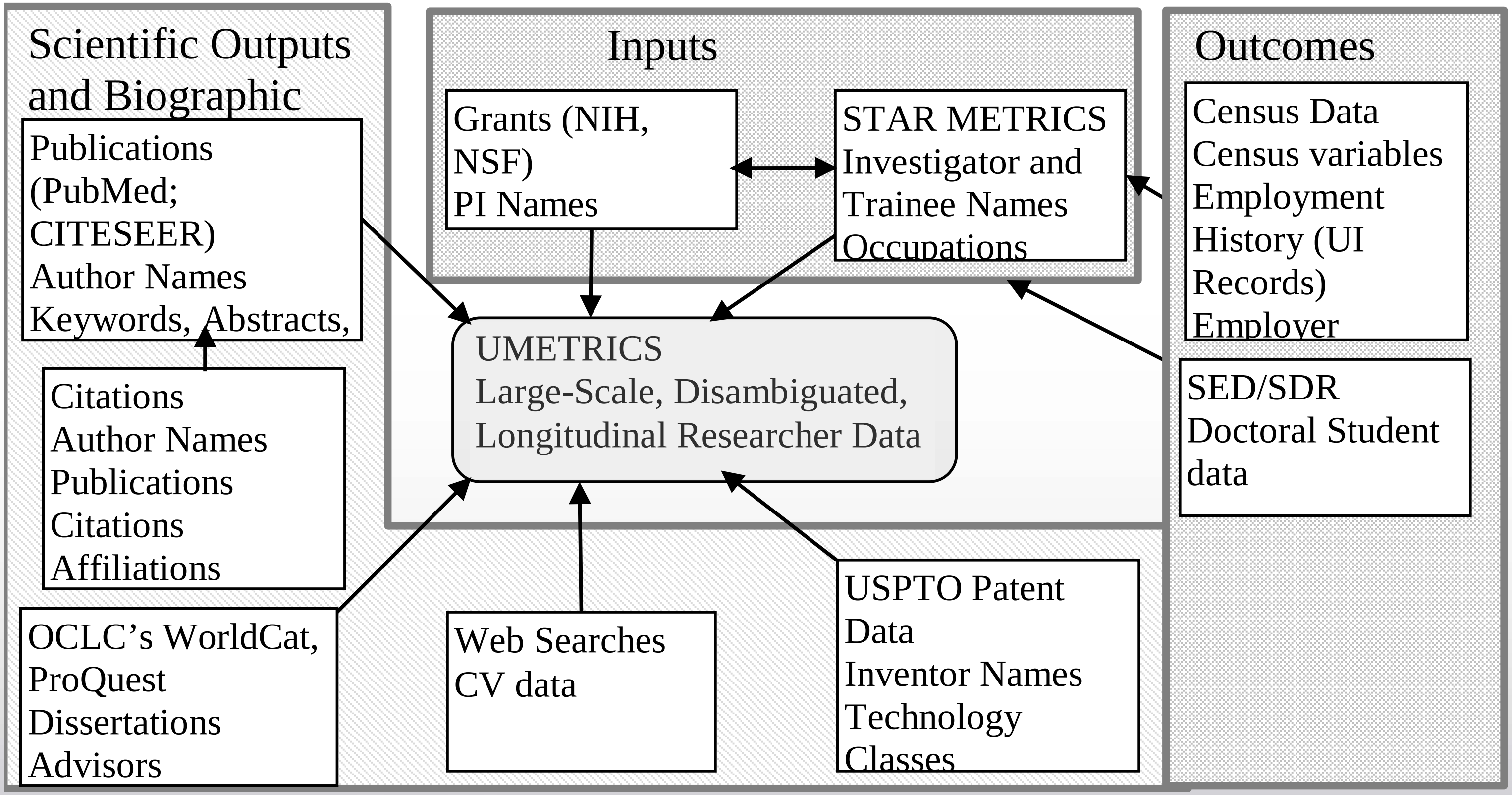
IMI Background

- Census Goals:
  - Improve measurement of small but important sector of the economy
  - Address data gaps in the measurement of innovation and relation to economic growth
  - Learn how to collaborate with data providers to deliver data products they value
  - Prototype project that can be scaled and extended to other sectors of the economy

IMI Background

- Innovative Aspects:
  - Collaboration with the University of Michigan’s Institute on Research in Innovation and Science (IRIS)
  - Experiment with utilizing “fat pipe” of data for a sector of the economy
    - The University data is complementary to business and household data at Census
  - Makes extensive use of skills our staff learned through the Big Data classes

The Emerging Large-Scale, Disambiguated, Longitudinal Researcher database



Scientific Outputs and Biographic

Publications (PubMed; CITESEER)
Author Names
Keywords, Abstracts,

Citations
Author Names
Publications
Citations
Affiliations

OCLC's WorldCat, ProQuest Dissertations
Advisors

Inputs

Grants (NIH, NSF)
PI Names

STAR METRICS
Investigator and Trainee Names
Occupations

UMETRICS
Large-Scale, Disambiguated, Longitudinal Researcher Data

Web Searches
CV data

USPTO Patent Data
Inventor Names
Technology Classes

Outcomes

Census Data
Census variables
Employment History (UI Records)
Employer

SED/SDR
Doctoral Student data

Establishment of new Institute

- Institute for Research on Innovation and Science (IRIS) founded 01/01/2015
  - Goal – leverage existing data to both serve university data and generate new research
  - Core facility at University of Michigan
  - 3 years seed funding for infrastructure from Sloan & Kauffman
- More efficient mode of data ingest for the Census Bureau
  - One MOU rather than N

(subset of) Preliminary Findings

- So far we’ve constructed basic indicators on:
  - Worker characteristics
  - Job placements (of students)
  - Vendor characteristics (including geographic patterns)
  - Startups
  - Patents
  - Trade

Job Placements - 1 Year After Leaving Institution

	Individuals on Grants		Proportion by Sector (6+Months)			Proportion by Sector (6+ Months & <50miles)		
Last Year	Overall	6 Months	Industry	Academia	Government	Industry	Academia	Government
2010	11,689	8,041	55.9%	36.0%	7.4%	22.9%	54.1%	19.6%
2011	19,049	13,562	63.1%	29.9%	6.3%	15.6%	49.5%	9.8%
2012	19,722	12,185	58.8%	34.4%	6.1%	20.9%	57.4%	20.0%

- The initial links suggest the main destination of grant recipients is Industry, followed by Academia
- Geographic matches very interesting, but can’t be shown for disclosure reasons

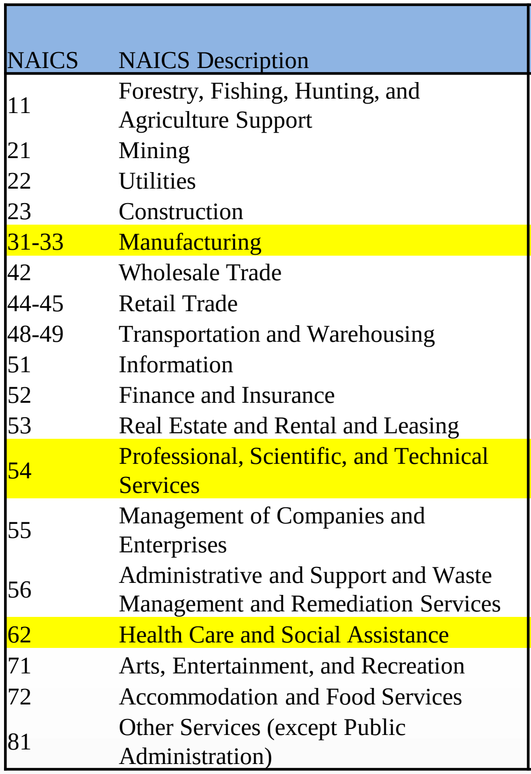
Job Placements - 1 Year After Leaving Institution

By Funding Source

	Individuals on Grants		Proportion by Sector (6+Months)			Proportion by Sector (6+ Months & <50miles)		
Funding Source	Overall	6 Months	Industry	Academia	Government	Industry	Academia	Government
NIH	17,336	13,684	61.5%	31.7%	6.1%	16.5%	49.5%	13.9%
NSF	7,118	4,784	56.5%	36.9%	6.0%	19.2%	49.9%	11.5%
Non-Federal	16,082	9,382	63.5%	28.2%	7.7%	18.7%	58.4%	16.7%
Dept of Education	2,852	1,383	49.0%	46.1%	4.3%	32.8%	69.0%	27.1%
Other	7,072	4,555	54.1%	38.7%	6.5%	25.0%	55.1%	21.5%

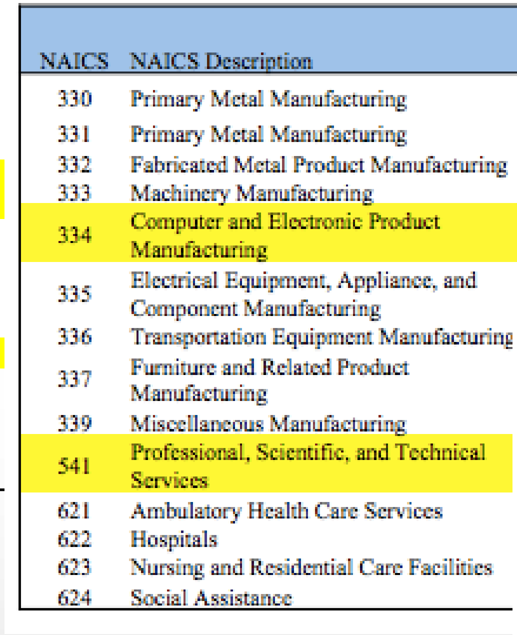
We can also break out Funding Source and Job Placements Relationship by School and Last Profession

2010 Cohort 2-digit NAICS



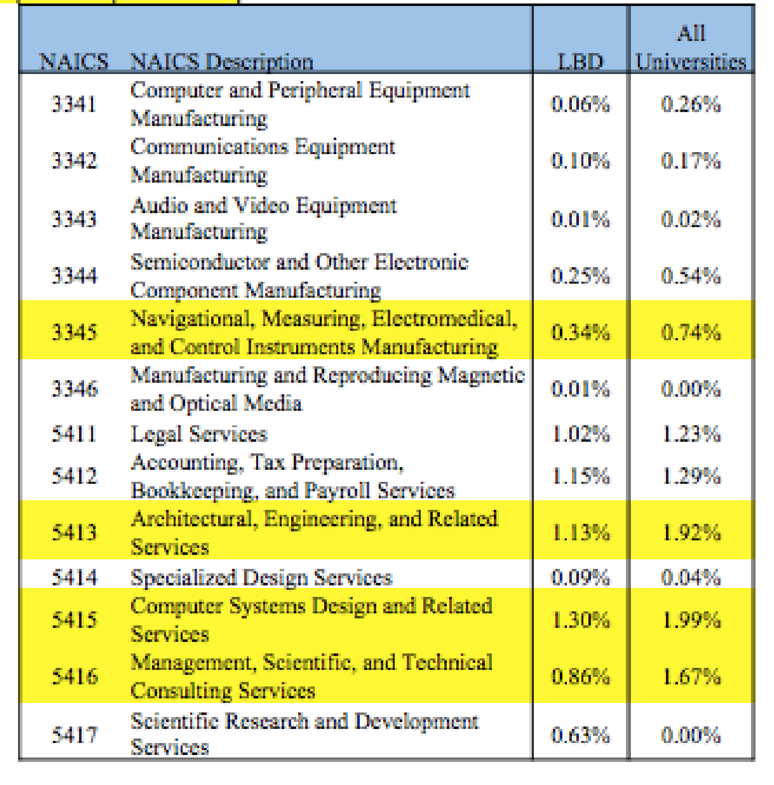
NAICS	NAICS Description	LBD	All Universities
11	Forestry, Fishing, Hunting, and Agriculture Support	1.12%	0.77%
21	Mining	0.59%	0.36%
22	Utilities	0.72%	0.32%
23	Construction	4.64%	2.63%
31-33	Manufacturing	9.75%	12.24%
42	Wholesale Trade		
44-45	Retail Trade		
48-49	Transportation and Warehousing		
51	Information		
52	Finance and Insurance		
53	Real Estate and Rental and Leasing		
54	Professional, Scientific, and Technical Services		
55	Management of Companies and Enterprises		
56	Administrative and Support and Waste Management and Remediation Services		
62	Health Care and Social Assistance		
71	Arts, Entertainment, and Recreation		
72	Accommodation and Food Services		
81	Other Services (except Public Administration)		

2010 Cohort 3-digit NAICS (Manufacturing)



NAICS	NAICS Description	LBD	All Universities
330	Primary Metal Manufacturing	0.00%	0.01%
331	Primary Metal Manufacturing	0.33%	0.28%
332	Fabricated Metal Product Manufacturing	1.18%	1.01%
333	Machinery Manufacturing	0.85%	1.38%
334	Computer and Electronic Product Manufacturing	0.78%	1.73%
335	Electrical Equipment, Appliance, and Component Manufacturing		
336	Transportation Equipment Manufacturing		
337	Furniture and Related Product Manufacturing		
339	Miscellaneous Manufacturing		
541	Professional, Scientific, and Technical Services		
621	Ambulatory Health Care Services		
622	Hospitals		
623	Nursing and Residential Care Facilities		
624	Social Assistance		

2010 Cohort 4-digit NAICS (Computer & Electronics Manufacturing)



NAICS	NAICS Description	LBD	All Universities
3341	Computer and Peripheral Equipment Manufacturing	0.06%	0.26%
3342	Communications Equipment Manufacturing	0.10%	0.17%
3343	Audio and Video Equipment Manufacturing	0.01%	0.02%
3344	Semiconductor and Other Electronic Component Manufacturing	0.25%	0.54%
3345	Navigational, Measuring, Electromedical, and Control Instruments Manufacturing	0.34%	0.74%
3346	Manufacturing and Reproducing Magnetic and Optical Media	0.01%	0.00%
5411	Legal Services	1.02%	1.23%
5412	Accounting, Tax Preparation, Bookkeeping, and Payroll Services	1.15%	1.29%
5413	Architectural, Engineering, and Related Services	1.13%	1.92%
5414	Specialized Design Services	0.09%	0.04%
5415	Computer Systems Design and Related Services	1.30%	1.99%
5416	Management, Scientific, and Technical Consulting Services	0.86%	1.67%
5417	Scientific Research and Development Services	0.63%	0.00%

Over/Under-Represented Industries

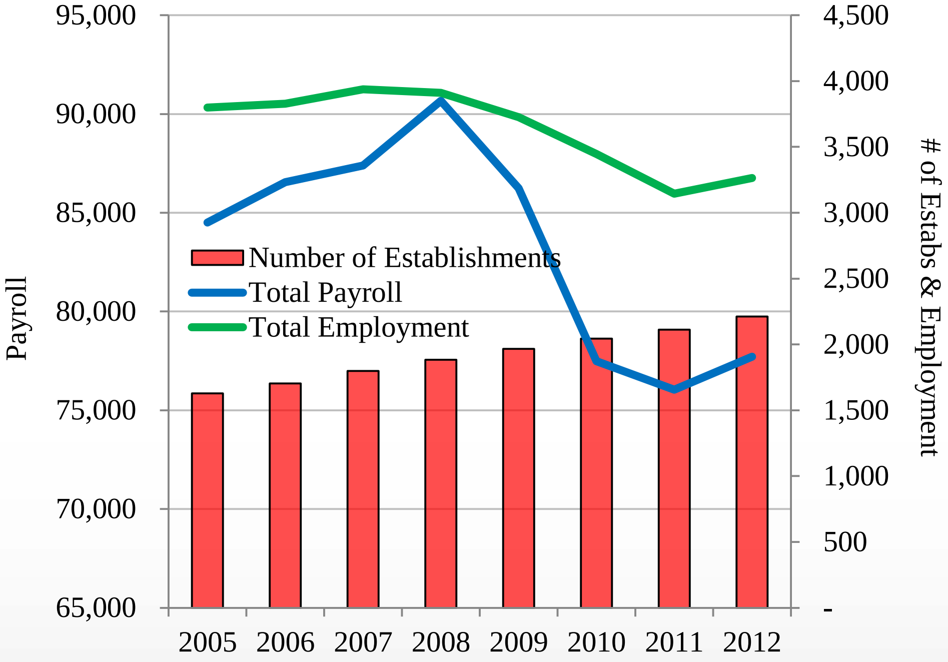
Most Overrepresented 4-digit NAICS

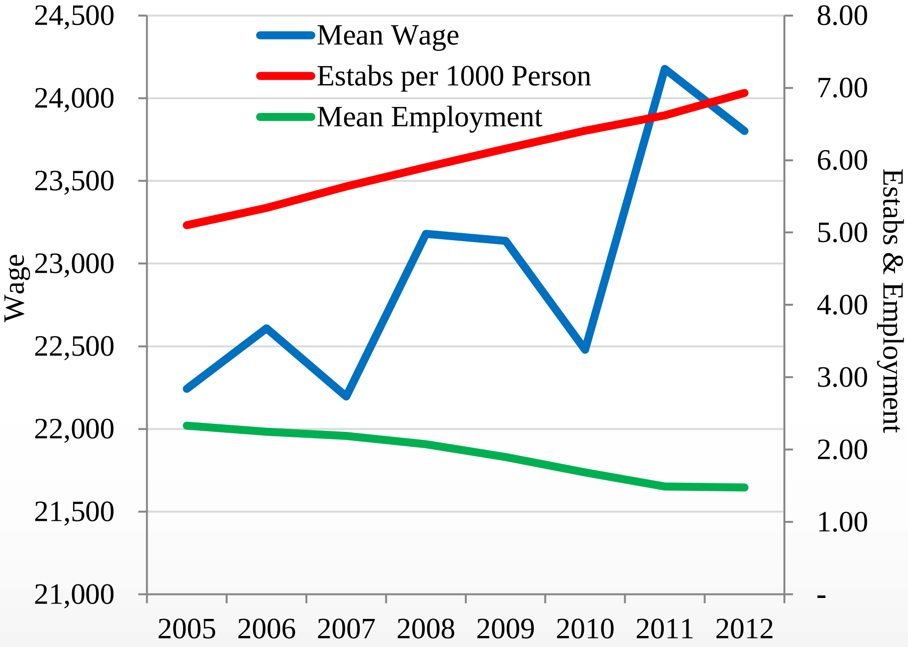
	NAICS	NAICS Description	U.S.	Univs.	Dif
1	5413	Architectural, Engineering, and Related Services	1.13%	4.34%	3.21%
2	5415	Computer Systems Design and Related Services	1.30%	3.97%	2.68%
3	5613	Employment Services	3.87%	6.26%	2.39%
4	5416	Management, Scientific, and Technical Consulting Services	0.86%	2.67%	1.82%
5	6221	General Medical and Surgical Hospitals	4.63%	5.96%	1.33%
6	4236	Electrical and Electronic Goods Merchant Wholesalers	0.43%	1.72%	1.28%
7	6214	Outpatient Care Centers	0.69%	1.82%	1.12%
8	8132	Grantmaking and Giving Services	0.17%	1.25%	1.08%
9	5112	Software Publishers	0.32%	1.35%	1.03%
10	5191	Other Information Services	0.23%	1.25%	1.02%

Most Underrepresented 4-digit NAICS

1	7222	Limited-Service Eating Places	3.63%	1.84%	-1.79%
2	4451	Grocery Stores	2.26%	0.69%	-1.58%
3	2382	Building Equipment Contractors	1.39%	0.27%	-1.12%
4	5221	Depository Credit Intermediation	1.80%	0.71%	-1.09%
5	4529	Other General Merchandise Stores	1.51%	0.44%	-1.07%
6	7211	Traveler Accommodation	1.66%	0.66%	-1.00%
7	7221	Full-Service Restaurants	4.03%	3.12%	-0.91%
8	8131	Religious Organizations	1.47%	0.56%	-0.90%
9	5617	Services to Buildings and Dwellings	1.46%	0.56%	-0.89%
10	6231	Nursing Care Facilities	1.46%	0.64%	-0.82%

Startup Business Dynamics (Matched through SS-4)





- Number of startups has been steadily increasing, although the cumulative size of these firms has been somewhat flat

Affiliated Projects

- 2020
  - Reengineering Address Canvassing
  - Planning and executing Non-Response Follow-up
  - Cost Reduction to Field Reengineering
- Retail Statistics

Retail Big Data Project Goal

To explore the use of “Big Data” to **supplement** existing monthly/annual retail surveys to fill in data gaps and increase relevance. Primary focus is to try to produce geographic level estimates more frequently than once every five years through the Economic Census.

The “Big” Data

- Posted initial RFI from 2/7/14 – 3/10/14
  - NAICS 518210; Data processing, hosting, and related services
- Posted 3 additional RFIs under different NAICS codes from 6/18/14 – 6/24/14
  - NAICS 522320; Financial transaction processing, reserve, and clearing house activities
  - NAICS 541910; Marketing research and public opinion polling
  - NAICS 522210; Credit card issuing
- Posted RFP from 7/15/14 – 8/15/14 under NAICS 541910
  - Received 2 submissions
- Awarded contract to NPD for two off-the-shelf datasets on 9/19/14
  - Automotive parts
  - Jewelry & watches
- Final datasets (2012-2014) received on 2/6/15

Retail Big Data Team Goal

To evaluate the data obtained from the NPD Group
to determine its usefulness in meeting the goal
of supplementing our retail statistics
with more frequent geographic level estimates.

About NPD

- NPD has agreements with approximately 900 retailers worldwide covering approximately 150,000 locations/stores and $400 billion in annual sales.
- Smaller businesses are generally not included
- Retailers provide aggregated (SKU-level) transaction data to NPD generally using a weekly feed (Sunday through Saturday) following the National Retail Federation reporting calendar
  - Store identifier/location
  - Item/Product code (e.g., SKU)
  - Dollar volume of sales
  - Units sold
  - Average price (calculated)
  - Flag distinguishing on-line from in-store sales
- NPD focus is on non-food and non-drug categories

Evaluation Plan

- Analyze NPD data to identify potential errors prior to use
  - Errors in geographic coding
  - Missing data
- Compare NPD data with Census Bureau estimates to obtain a rough assessment of coverage and to determine if the NPD data could serve as a potentially informative predictor
  - Aggregate levels (monthly or yearly totals)
  - Period-to-period changes (e.g., current-to-prior month, current-to-prior quarter, current month to same month a year ago)
- Census Bureau Data Available for Comparisons
  - Monthly Retail Trade Survey
  - Annual Retail Trade Survey
  - 2012 Economic Census
  - Business Register

Current Status

- Analyzed NPD data to identify potential errors
- Compared levels and period-to-period changes of NPD data to sales estimates derived from the Monthly Retail Trade Survey
- Started extracting 2012 Economic Census data for companies included in the NPD totals
- Next Steps: Finish gathering Census Bureau company and establishment data and complete additional comparisons using these data. Start summarizing findings into a draft report.

Other Possibilities

- Explore Feasibility of Obtaining Data Feeds Directly from Retail Companies
  - Agreements with individual companies
  - Access through 3rd party such as NPD, Nielsen, or IRI

- Benefits
  - May reduce reporting burden on companies
  - Would allow us to obtain more detailed data more frequently
  - Leveraging a 3rd party could help with standardized formats

- Test with a few companies in 2017 Economic Census

- Obtain store level data from credit card transactions (Mastercard, 3rd party processors, banks, etc.)

Background Slides

Innovation and Research

- Goal of research project/firm: to create and transmit scientific ideas and push for their adoption (by other scientists, policy-makers or businesses)
- Behavioral Framework; Ideas are transmitted by workers in a variety of potentially measurable ways
- Behavioral Framework: Social networks/collaboration are a major vehicle whereby ideas are transmitted

Empirical Measurement

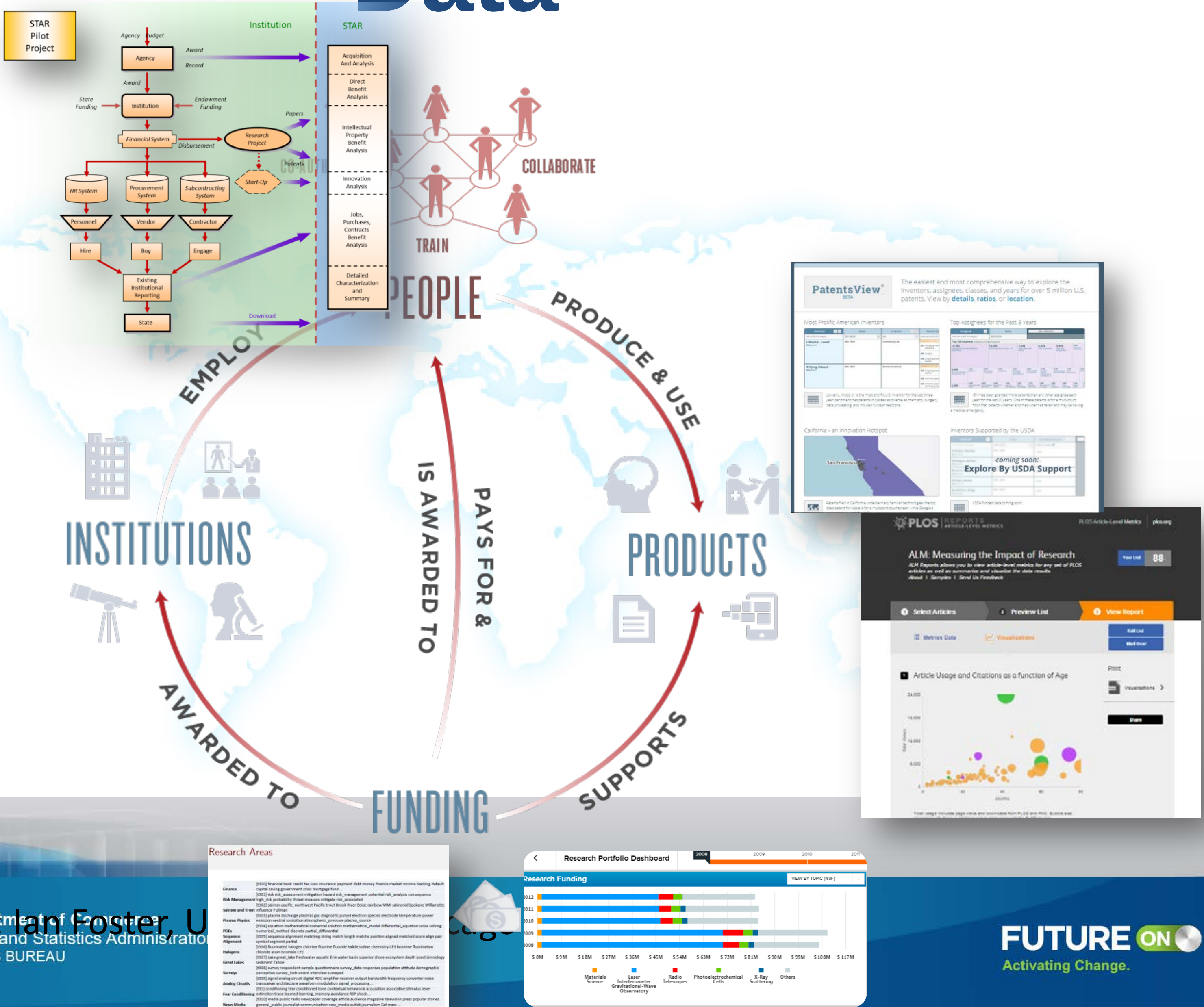
Transmission of ideas can occur through:

- People employed doing research (measured with grants)
  - Placement of individuals
  - Start up of businesses
  - Through social networks
- Purchases of equipment and services
  - Consumer led innovation
  - Development of comparative advantage
  - Economies of scale

Sample Questions

- Role of social and physical distance; basic/applied research; gender/ethnicity; teams….

The Empirical Framework: Big Data



PatentsView BETA

The easiest and most comprehensive way to explore the inventors, assignees, classes, and years for over 5 million U.S. patents. View by details, ratios, or location.

Most Prolific American Inventors

Lowell L. Wood Jr. is the most prolific U.S. inventor for the last three-year period and has patents in classes as diverse as chemistry, surgery, data processing, and induced nuclear reactions.

Top Assignees for the Past 3 Years

IBM has been granted more patents than any other assignee each year for the last 20 years. One of these patents is for a multi-touch floor that detects whether a homeowner has fallen and may be having a medical emergency.

California - an Innovation Hotspot

San Francisco

Patents filed in California underlie many familiar technologies: the top cited patent for Apple is for a multipoint touchscreen, while Google's

Inventors Supported by the USDA

coming soon:
Explore By USDA Support

USDA funded data coming soon...

PLOS | REPORTS ARTICLE-LEVEL METRICS

PLOS Article-Level Metrics plos.org

ALM: Measuring the Impact of Research

ALM Reports allows you to view article-level metrics for any set of PLOS articles as well as summarize and visualize the data results.

About | Samples | Send Us Feedback

Your List 88

Select Articles · Preview List · View Report

Metrics Data · Visualizations

Article Usage and Citations as a function of Age

Print

Research Areas

Finance	[t300] financial bank credit tax loan insurance payment debt money finance market income banking default capital saving government crisis mortgage fund ...
Risk Management	[t301] risk risk_assessment mitigation hazard risk_management potential risk_analysis consequence high_risk probability threat measure mitigate risk_associated
Salmon and Trout	[t302] salmon pacific_northwest Pacific trout Brook River Boise rainbow MMI salmonid Spokane Willamette influence Fullman
Plasma Physics	[t303] plasma discharge plasmas gas diagnostic pulsed electron species electrode temperature power emission neutral ionization atmospheric_pressure plasma_source
PDEs	[t304] equation mathematical numerical solution mathematical_model differential_equation solve solving numerical_method discrete partial_differential
Sequence Alignment	[t305] sequence alignment matching string match length matche position aligned matched score align pair symbol segment partial
Halogens	[t306] fluorinated halogen chlorine fluorine fluoride halide iodine chemistry CF3 bromine fluorination chloride atom bromide CF2
Great Lakes	[t307] Lake great_lake freshwater aquatic Erie water basin superior shore ecosystem depth pond Limnology sediment Tahoe
Surveys	[t308] survey respondent sample questionnaire survey_data responses population attitude demographic perception survey_instrument interview surveyed
Analog Circuits	[t309] signal analog circuit digital ADC amplifier receiver output bandwidth frequency converter noise transceiver architecture waveform modulation signal_processing...
Fear Conditioning	[t31] conditioning fear conditioned tone contextual behavioral acquisition associative stimulus lever extinction trace learned learning_memory avoidance RSP shock...
News Media	[t310] media public radio newspaper coverage article audience magazine television press popular stories general_public journalist communication new_media outlet journalism Caf mass ...

Research Portfolio Dashboard

2008 2009 2010 2011

Research Funding

VIEW BY TOPIC (NSF)

Source: Ian Foster, U[illegible]cago

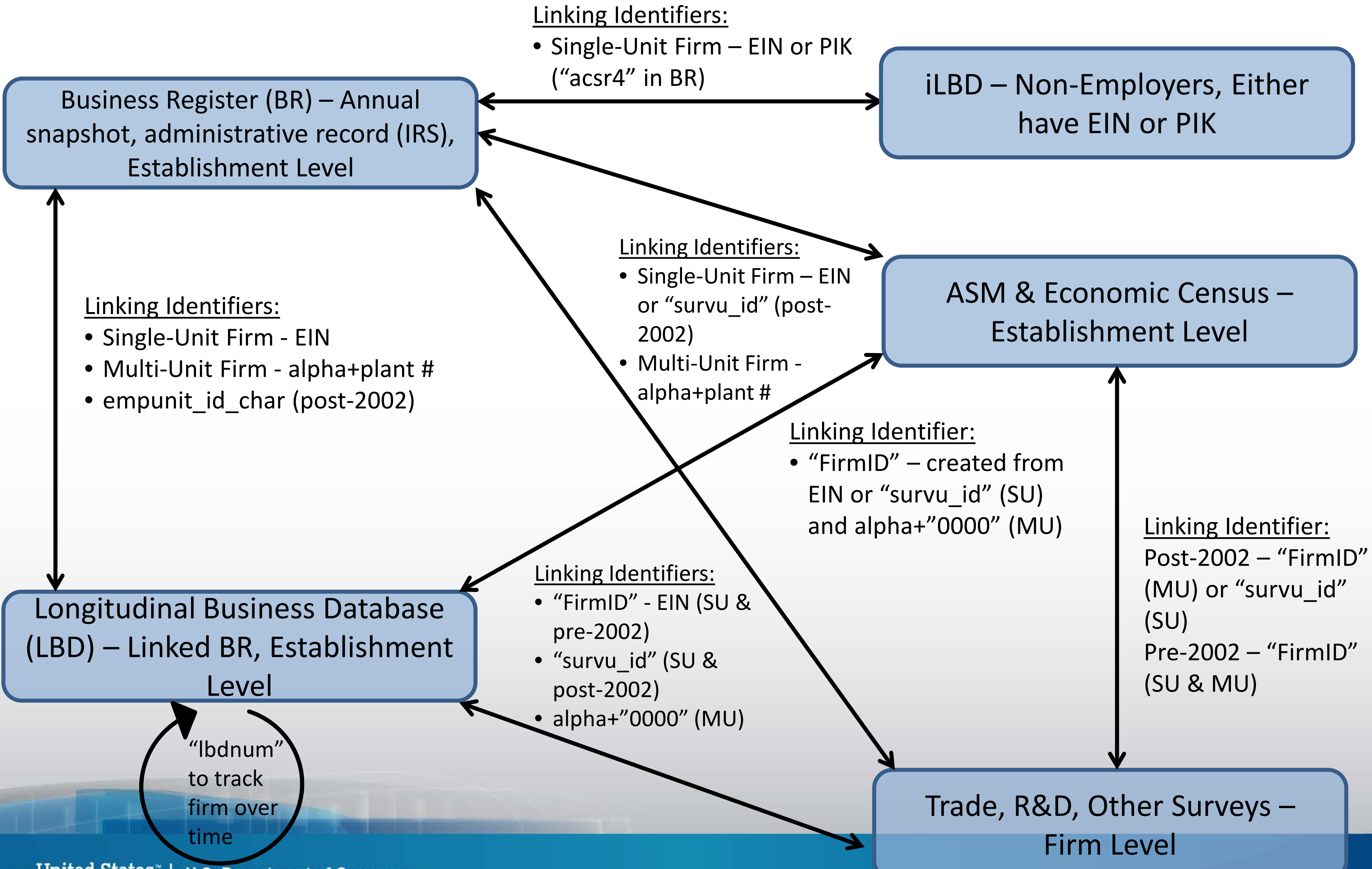
Data Description

- Business Register (BR)
  - Universe of U.S. non-agricultural businesses and the source of data from which all other economic data are ultimately created
  - Key data provided: Industry Classification (NAICS), Geographic data, Employment, Payroll, EIN Codes, Available from 2002-2012
- Longitudinal Business Database (LBD)
  - Universe of employer businesses, unique establishments, the LBD covers all industries and all U.S. States linked over time
  - Key data provided: Industry Classification (NAICS), Geographic data, Employment, Payroll, Firm Age, Available from 2002-2012
- Integrated Longitudinal Business Database (iLBD)
  - Universe of non-employer businesses with links to employer universe
  - iLBD records are identified by either PIKs or EINS, 85-88% are PIKs and 12-15% are EINS
  - Key data provided: Industry Classification, Gross Receipts, Geographic data, Available from 2002-2010
- Longitudinal Employer-Household Dynamics (LEHD)
  - Employee-Employer linked dataset
  - Key data provided: EIN-Geocode Linkage, Wage Data, Available from 2002-2010
- W2 Data
  - Key data provided: PIK, Wage Data, Available from 2005-2012

Data Description

- Economic Census (every 5-years ending in "2" or "7")
  - Comprehensive survey of 9 separate sectors of economy including Manufacturers, Mining, Utilities, Wholesale, Retail, Agriculture, Transportation, Services and Finance
  - More detailed firm performance data: firm expenditures, receipts, production hours, etc...
- Other surveys
  - Annual Survey of Manufacturers (between 50,000-60,000 firm-level observations)
  - R&D Surveys (annual, between 25,000-30,000 firm-level observations)
  - Foreign Trade Transactions (transaction-level data on imports and exports, firm-level identifiers)
  - Mining and energy use surveys (firm-level identifiers, every 5-years)
  - Commodity flow survey – movement of goods (every 5-years)
- STAR METRICS
  - Transaction-level data on grant recipients at major universities
    - Vendors
    - Personnel (Faculty, Graduate Students, Technicians, etc…)
  - Award information – topic, funding agency, award amount, duration
  - Continuously updated (real time)

Construction



Business Register (BR) – Annual snapshot, administrative record (IRS), Establishment Level

iLBD – Non-Employers, Either have EIN or PIK

ASM & Economic Census – Establishment Level

Longitudinal Business Database (LBD) – Linked BR, Establishment Level

Trade, R&D, Other Surveys – Firm Level

Linking Identifiers:
- Single-Unit Firm – EIN or PIK ("acsr4" in BR)

Linking Identifiers:
- Single-Unit Firm - EIN
- Multi-Unit Firm - alpha+plant #
- empunit_id_char (post-2002)

Linking Identifiers:
- Single-Unit Firm – EIN or "survu_id" (post-2002)
- Multi-Unit Firm - alpha+plant #

Linking Identifier:
- "FirmID" – created from EIN or "survu_id" (SU) and alpha+"0000" (MU)

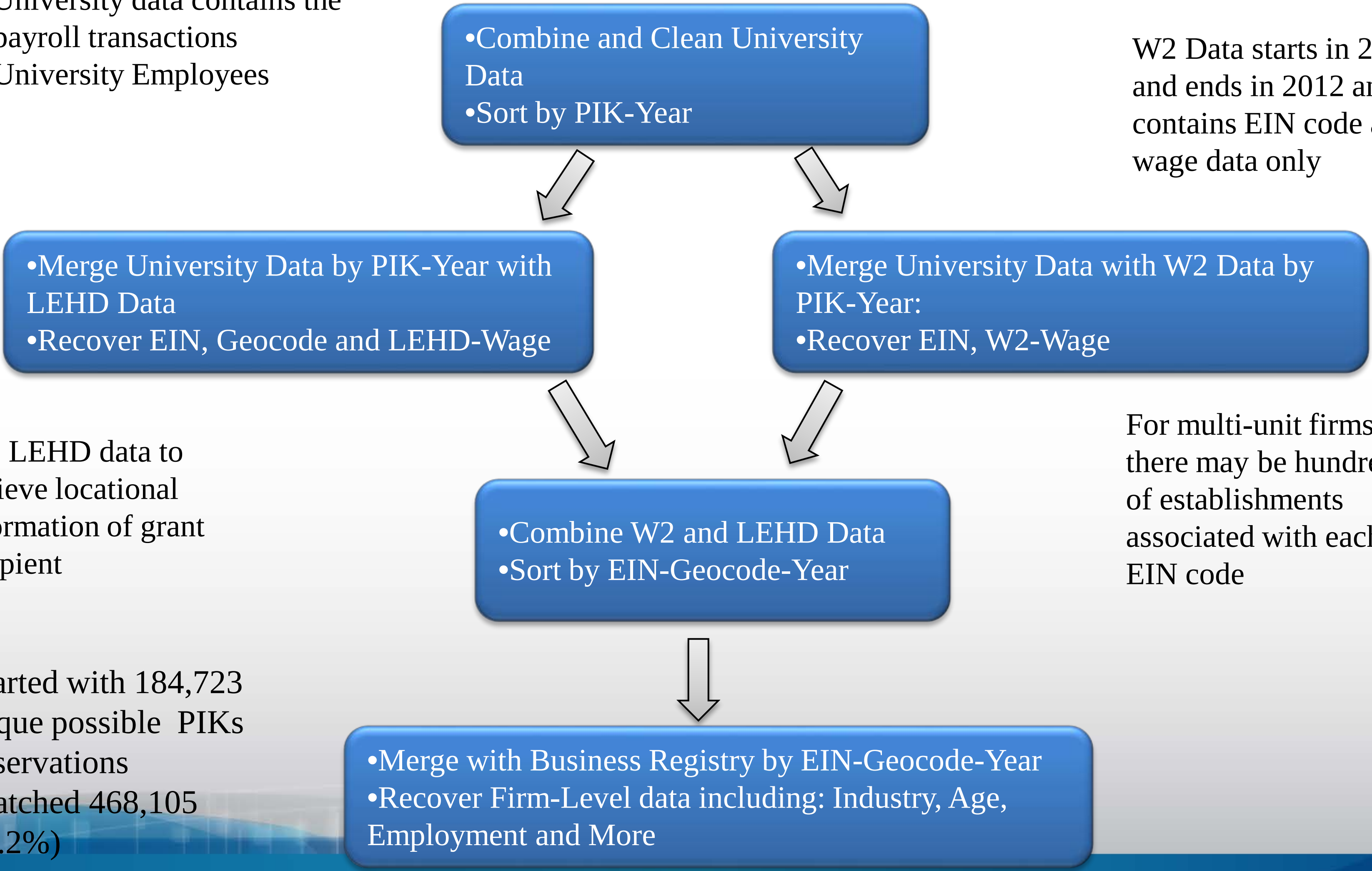
Linking Identifier:
Post-2002 – "FirmID" (MU) or "survu_id" (SU)
Pre-2002 – "FirmID" (SU & MU)

Linking Identifiers:
- "FirmID" - EIN (SU & pre-2002)
- "survu_id" (SU & post-2002)
- alpha+"0000" (MU)

"lbdnum" to track firm over time

Matching Process

University data contains the payroll transactions University Employees



- Combine and Clean University Data
- Sort by PIK-Year

W2 Data starts in 2005 and ends in 2012 and contains EIN code and wage data only

- Merge University Data by PIK-Year with LEHD Data
- Recover EIN, Geocode and LEHD-Wage

- Merge University Data with W2 Data by PIK-Year:
- Recover EIN, W2-Wage

Use LEHD data to retrieve locational information of grant recipient

- Combine W2 and LEHD Data
- Sort by EIN-Geocode-Year

For multi-unit firms, there may be hundreds of establishments associated with each EIN code

- Started with 184,723 unique possible PIKs Observations
- Matched 468,105 (98.2%)

- Merge with Business Registry by EIN-Geocode-Year
- Recover Firm-Level data including: Industry, Age, Employment and More